APPLICATION OF NEW TECHNOLOGIES TO DTIC DOCUMENT PROCESSING

August 1987

Approved for Public Release; Distribution Unlimited

AEROSPACE STRUCTURES
INFORMATION AND ANALYSIS CENTER

OPERATED FOR THE FLIGHT DYNAMICS LABORATORY
BY ANAMET LABORATORIES, INC.

SECURITY CLASSIFICATION OF THIS PAGE

AD-A189778

(2)

DTIC ELECTE MAR 07 1988 S H

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION: UNCLASSIFIED	1b. RESTRICTIVE MARKINGS: none
2a. SECURITY CLASSIFICATION AUTHORITY: N/A	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for Public Release; Distribution Unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE: N/A	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): 887.1A	5. MONITORING ORGANIZATION REPORT NUMBER(S): N/A
6a. NAME OF PERFORMING ORGANIZATION: Anamet Laboratories, Inc. / 6b. OFFICE SYMBOL (If applicable):	7a. NAME OF MONITORING ORGANIZATION: Air Force Wright-Aeronautical Laboratories Air Force Systems Command (AFWAL/FIBRA)
6c. ADDRESS (City, State, and ZIP Code): 3400 Investment Blvd. Hayward, CA 94545-3811	7b. ADDRESS (City, State, and ZIP Code): Wright-Patterson AFB, OH 45433
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: Flight Dynamics Laboratory / 8b. OFFICE SYMBOL (If applicable): AFWAL/FIBRA	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: F33615-84-C-3216
8c. ADDRESS (City, State, and ZIP Code): Air Force Wright-Aeronautical Laboratories, Air Force Systems Command, Wright-Patterson AFB, OH 45433	10. SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO. 62201F; PROJECT NO. 2401; TASK NO 02; WORK UNIT ACCESSION NO. 65

11. TITLE (Include Security Classification)

Application of New Technologies to DTIC Document Processing

12. PERSONAL AUTHOR(S)

Harris, Steven G.; Citerley, R. L. ; Cahn, David F.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM Jun 87 TO Aug 1987	August 1987	34

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Optical Character Recognition, Document Management, Database Management, Cataloging
05	02		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The use of integrated optical character recognition (OCR) and database management technology to improve DTIC document input processing is examined. Significant near-term improvements in efficiency can be realized using commercially available components fused into an integrated system. An approach and system architecture are defined that will permit a staged implementation of this technology within the framework of the current DTIC work flow. While the emphasis of this report is on reducing labor-intensive manual keystroking operations presently in use, the proposed system provides an open ended approach which will interface easily with both existing and future DTIC operations. An in-depth review of current DTIC document processing was used to guide the definition of the pilot system architecture developed in this study, and the results of that review are presented and discussed.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL: Duane Veley	22b. TELEPHONE (Include Area Code): (513) 255-7191 / 22c. OFFICE SYMBOL: AFWAL/FIBRA

DD Form 1473, JUN 86 — Previous editions are obsolete. — SECURITY CLASSIFICATION OF THIS PAGE

88 3 5 076

FOREWORD

This report summarizes the results of a requirements study performed by Anamet Laboratories for the Defense Technical Information Center, Office of Information Systems and Technology. The study defines the feasibility and approach for implementing a pilot system that integrates optical character recognition and database management technology to improve document input processing within the DTIC workflow environment.

This work was performed by Anamet Laboratories, under Task 4.2-28 for the Aerospace Structures Information and Analysis Center (ASIAC). Anamet operates ASIAC for the Flight Dynamics Laboratory at the Wright-Patterson Air Force Base under contract number F33615-84-C-3216. Database Applications, Inc. was a consultant to Anamet on this effort.

Approved by:

M. Dilip Bhandarkar
ASIAC Program Manager

Prepared by:

Steven G. Harris
Senior Engineer
Anamet Laboratories, Inc.

Richard L. Citerley
Vice President
Anamet Laboratories, Inc.

David F. Cahn
President
Database Applications, Inc.

Contents

List of Figures

EXECUTIVE SUMMARY

Recent technological innovations hold considerable promise for the Defense Technical Information Center's document database publication activities. Solid state image scanning, image to digital text conversion, optical disk mass storage, and high-speed electronic printing can revolutionize DTIC's present labor-intensive document handling procedures and can improve both the quality and production time for DTIC products.

Each of these technologies has been developing independently and has recently matured to the point at which full-scale production use may be viable. Production operation involves meshing of the components into a fully integrated and coordinated system. To accomplish this integration, it is necessary to accommodate the individual limitations of the state-of-the-art components which are used, and to superimpose an executive control function to coordinate information flow among them and to allow them to work in unison.

DTIC has tasked Anamet Laboratories to review DTIC's current operations and to provide a realistic assessment of the potential improvements that these new technologies can provide. The outcome of that review is contained in this document. It provides an approach and system architecture which will permit a staged implementation of this technology within the framework of the current DTIC work flow. While the emphasis in this effort has been on reducing labor-intensive manual keystroking operations presently in use, the proposed system provides an open ended approach that will interface easily with both existing and future DTIC operations.

1 INTRODUCTION

This report summarizes the results of a requirements study performed by Anamet Laboratories for the Defense Technical Information Center (DTIC) under Task 4.2–28 of the Aerospace Structures Information and Analysis Center (ASIAC). The study defines the requirements for a pilot system that integrates Optical Character Recognition (OCR) and database management technology to provide a cost–effective means of bringing paper–based documents into an online database. This online information is then subsequently loaded into the Defense RDT&E On Line System (DROLS). The primary initial goal of the pilot system is to improve DTIC's efficiency in the labor–intensive keystroking operations that are presently used to transform paper–based information into online data. At the same time, the pilot system provides an extendable baseline for meeting the longer term goals DTIC has envisioned for its Electronic Document System (EDS) [1].

In defining requirements for the pilot system, an important initial step was to study the current DTIC document processing work flow. This study provided a guide for determining the optimum pilot system implementation approach, so as to maximize benefits while minimizing disruption to existing procedures. Anamet's previous work on related Air Force document management problems was also brought to bear in defining both the pilot system requirements and an initial system for demonstration and evaluation.

The following section describes how the generic components of a document acquisition system can be tailored to meet specific DTIC requirements. The resulting pilot system architecture reflects the findings of the work flow study, input from DTIC personnel, Anamet's experience in related Air Force projects, and the realistic constraints of the state–of–the–art technology.

2 DOCUMENT ACQUISITION SYSTEM TECHNOLOGY

Recent technological advances now make it feasible to assemble a comprehensive system to bring offline documents into an online database. This document acquisition system can be assembled primarily using commercially available components. Some of the rapidly maturing technologies that can contribute to a document acquisition system include:

- Image scanners
- Optical Character Recognition (OCR) devices
- Image processing software and hardware
- Database management software

- **Local Area Networks (LAN)**
- Optical Storage Devices

These components are available from a variety of manufacturers, and their costs vary across a wide spectrum depending on the level of capability required in the assembled system.

2.1 Generic System Components

A generic configuration of a document acquisition system can be considered to consist of the five basic components shown in Figure 1. The input component provides the means for documents to enter the system; for DTIC, these documents are paper–based forms, and they enter the system as page images. The conversion component provides the means to transform these page images to a computer–readable form: ASCII text. A verification, or proof editing, pass on the text that results from this automated process forms an important part of the total conversion process. The computer–based documents that result from the input and conversion processes must then be loaded into the database and stored online.

For the initial stages of the DTIC pilot system, captured documents will be "dynamic" in nature, changing as they pass through the DTIC work flow, and magnetic disk is the preferred storage medium. For longer term DTIC applications which deal with "static" unchanging documents, optical disk storage is a feasible alternative. The retrieval component of the system provides the means to locate and retrieve any document, using a variety of methods. Finally, the overall flow through the input, conversion, storage and retrieval components is controlled from a master document control subsystem. This control component provides the central point of access for all system users, tracks document progress as it flows through the system, and screens user access to the evolving central library.

2.2 Tailoring the Technology to DTIC Requirements

The generic components described above must be assembled in a system that is tailored to the document input flow at DTIC. It is important to note, however, that most of the individual hardware and software components that comprise the system are available off–the–shelf. The manner in which these components are assembled to form an integrated system provides the customized solution to meet specific document processing requirements.

DTIC's primary need is to reduce labor–intensive keystroking operations presently used to process forms–based paper documents. Within this scope, the primary form being addressed is the DD1473. The DD1473 provides catalog information on each Technical Report (TR), and it is supplied to DTIC as a bound page inside each report. In addition, DTIC desires the system to be capable of handling other relevant forms–based information, including the DTIC271 (Independent R & D Data Sheet)

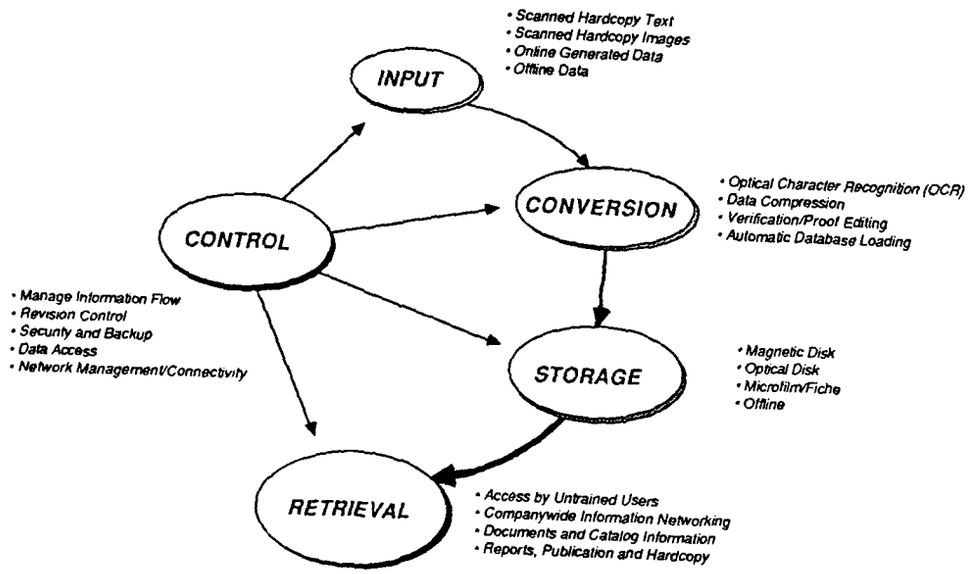


Figure 1: Document Acquisition System Components

and DD1498 (Research and Technology Work Unit Summary) forms, and the Program Element Descriptive Summary (PEDS). Note that, at present, all information contained on the DD1473 and other forms is manually entered into DTIC databases. This keystroked data goes through extensive checking and modification as it flows through the document input process.

DTIC's desire to minimize disruption to the present work flow is a key driver in customizing the system configuration. The intent is to apply the new technology selectively to key DTIC operations, and to gradually migrate more staff and operations online over time. To minimize the disruption, an in-depth knowledge of the DTIC document processing is required. Candidate configurations are examined to determine their adequacy to mesh with all of the detailed steps performed by DTIC personnel in transposing forms-based information into accurate online data for entry into the Technical Report (TR) database. Some of the parameters that enter into the evaluation of candidate configurations include:

1. Document throughput, present versus projected
2. Document volume, per DTIC document processing cycle
3. Skill level of personnel available to perform various functions
4. Separation of data entry functions from cognitive indexing functions
5. System supportability and extendability

These parameters are then balanced with the realistic limitations of the hardware and software components, as well as system cost, to arrive at the proposed configuration.

2.3 Previous Work On MAIRS

As a previous task under its ASIAC contract, Anamet Laboratories developed a prototype "assembly line" system for bringing paper-based Military Standards into an online searchable database. This effort was called MAIRS (MIL-STD Automated Indexing and Retrieval System) [2,3,4]. A baseline system architecture was defined to integrate OCR and database management technologies, and a prototype system was assembled to demonstrate the basic system concepts.

Concepts which were developed under the MAIRS effort provide a foundation for the DTIC pilot system. In addition, the knowledge gained in development work for MAIRS provides an invaluable step up the "learning curve" in providing integrated solutions to document management problems. For example, when MAIRS was begun, only one OCR device was available that had the necessary flexibility for incorporation into a fully integrated document acquisition system, and Anamet Laboratories acted as a "beta" test site in adapting it to MAIRS. In addition, extensive evaluations were performed of database/document management systems and optical disk storage systems in support of MAIRS.

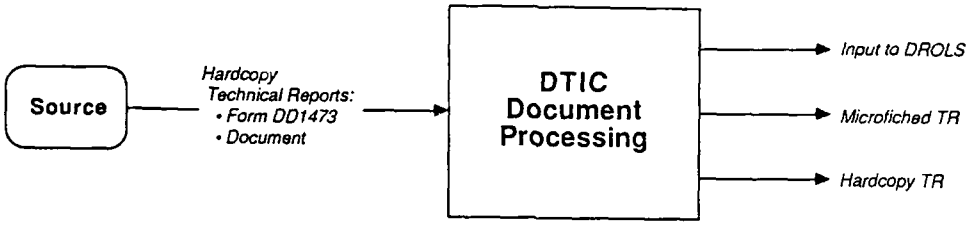


Figure 2: Overview of DTIC Document Processing

The MAIRS project has therefore provided Anamet and the DoD with a prototype system for document acquisition, as well as the ability to define clearly what is and is not feasible in a variety of rapidly maturing technologies. This experience provides a starting point for the DTIC pilot system and an unbiased knowledge base upon which to make informed implementation decisions.

3 DTIC PILOT SYSTEM DESCRIPTION

The DTIC pilot system is designed to operate in the production DTIC environment and should provide a significant increase in efficiency by reducing manual keystroking operations. Its form and functionality are guided by the complex document processing steps used by DTIC, as well as the limitations of the state–of–the–art technology used in the system. The pilot system will be implemented in a staged manner, to provide for DTIC evaluation of its performance, and to minimize disruption to the current DTIC work flow.

Figure 2 provides a very broad overview of DTIC processing of Technical Reports. The pilot system will address the part of this document processing which transforms information delivered to DTIC on a DD1473 form into online data in preparation for entry into the TR database. The Stage I pilot system focuses on data entry operations, and the Stage II system extends to increase efficiency in the cognitive indexing and administrative functions through increased access to online data.

3.1 DTIC Pilot System Functionality

Figure 3 illustrates how the pilot system functions. Paper–based forms are brought online initially as page images, using a commercially available image scanner. Such devices allow page scanning of bound material (much like a photocopy machine) so that the DD1473 need not be separated from the TR. An adjustable set of "zones" is

used to identify areas on the page image that correspond to database fields in the central library. The image in each zone is then passed to an OCR device that converts the image to digital (ASCII) text. All character recognition is performed in a background mode, and when it is complete for a particular document, the central controller makes the "raw" character interpretation available to operators for verification.

The verification process is a key element in the flow. Because of smudges, distortion or poor document quality, the OCR device cannot always properly interpret as characters the image which it sees. A verification, or proof editing, pass is required to correct any deficiencies in the OCR interpretation. The verification is performed online. The software automatically positions the cursor at each character in the converted text for which conversion was less than certain, and shows the operator an image of the original page in the immediate vicinity. Operators can overtype or take other corrective action as desired, pressing a single key to advance to the next area that needs attention when they are through. Intervening portions of the form, however lengthy, are skipped. Operators only make corrections; the system flags and searches for the errors automatically. The verification operation represents a significantly less labor intensive process than does manual keystroke entry. The verified document is then passed back to the central library for storage.

At this point, all the information available from the DD1473 is available for loading into the database. The information has been verified by operators who would previously have performed keystroking. A mapping between zones on the form and fields on electronic cards in the central library database permits the verified data to be loaded automatically, as shown in Figure 4. No manual database entry is required. Additional modifications to the document, such as may be required for subsequent DTIC processing, take place after the OCR verification pass, and before final release of the information for inclusion in the TR database.

3.2 Staged Implementation Approach

The DTIC pilot system will be implemented in a staged manner, each stage building on the work of previous stages. An enhanced version of the prototype system developed for the MAIRS effort is considered to be Stage 0, since it provides a demonstration of many of the concepts necessary for the full pilot. The Stage I system will concentrate on the immediate task of reducing the volume of keystroke operations presently performed by DTIC. The Stage I system is designed to maximize the impact on DTIC's efficiency in the input process, while minimizing the disruption to current work flow. Follow–on stages are envisioned to make use of existing software capability to more fully automate the input process, to provide administrative reports and control over this process, and to expand access to the system.

3.2.1 Stage 0 System

The demonstration system developed under the MAIRS effort (see Section 2.3) was enhanced to apply more directly to the DTIC effort under ASIAC Task 4.2–28. The

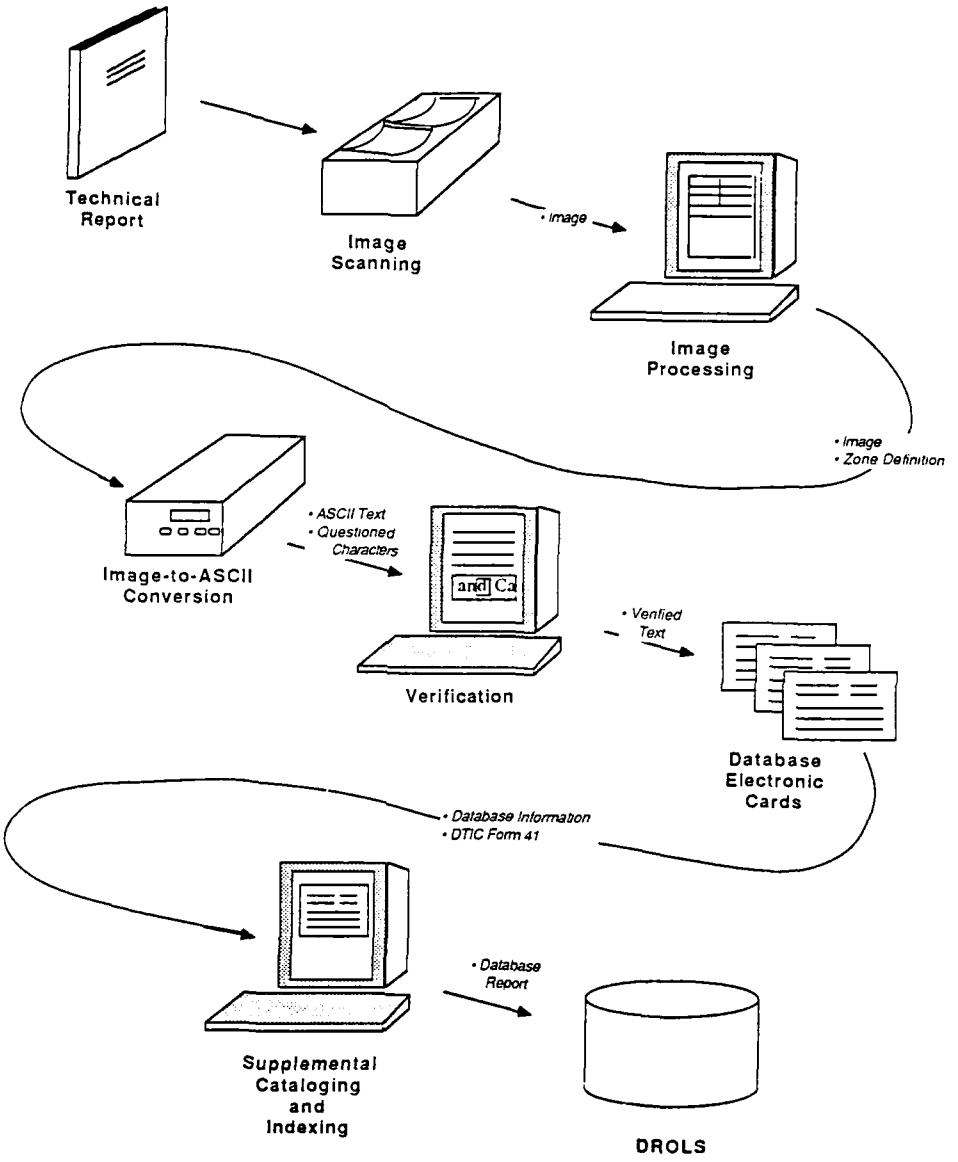


Figure 3: Pilot System Document Processing Flow

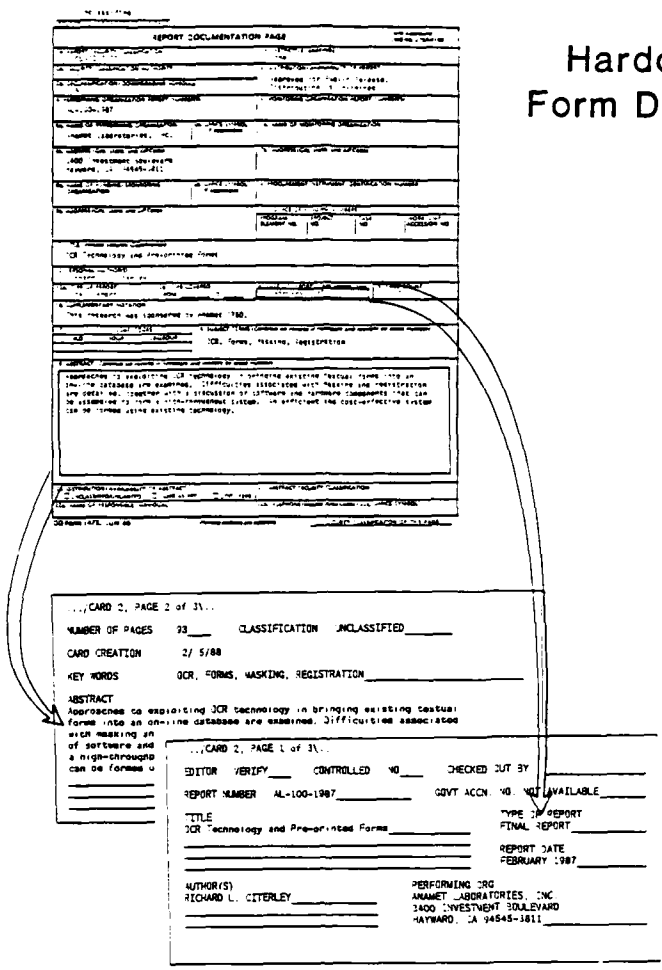


Figure 4: Auto–loading of Database Fields from Forms

resulting configuration is termed the Stage 0 DTIC pilot system, and it was successfully demonstrated at DTIC during July 1987.

The Stage 0 system is entirely IBM/PC–based, and consists of a single workstation to perform scanning, OCR conversion, verification editing, database loading, and retrieval functions. Custom user–designed electronic card forms were demonstrated, matching the different types of hardcopy forms which were scanned. Automated loading of the electronic cards from hardcopy forms was demonstrated for the DD1473, DTIC271 and DD1498. Many of these operations will take place at physically separate workstations in the Stage I and subsequent pilot systems. The Stage 0 system provides a real working prototype and serves to demonstrate some of the key concepts involved in subsequent pilot systems, including:

1. Smooth integration of OCR hardware, control software, and database software.
2. Direct database loading of forms–based information.
3. An extendable workstation concept.
4. Verification (proof editing) of scanned documents, tuned to the type of database field being examined.
5. Significant throughput increase over manual keystroke operations.

3.2.2 Stage I System

In Stage I, while some technology and work flow issues remain to be resolved, it is recommended that most of the paper–based work flow currently in use at DTIC be retained. The key difference in operation is that data are actually collected online via scanning and OCR equipment, then accumulated in a central database. The paper documents currently used for keyword assignment and other document processing activities will still be used, but will be generated from this database as output reports. The intent is to retain the work flow of the majority of DTIC personnel exactly as at present, while integrating the system components and bringing them up to production levels. In Stage I, only the personnel currently involved with keystroking and direct inquiry against the Current File (the file of documents being processed in the current cycle) will have direct contact with the system. Others who are currently using paper will continue to do so.

3.2.3 Stage II System

In Stage II, DTIC personnel who are currently operating on paper will be brought online as appropriate, on a phased basis. Routing of paper–based information will be supplanted by direct access to the online data. Various offline data structures against which DTIC personnel crosscheck the documents being processed will be brought online as well. The online generation of MiniMAD data (the output product of Stage I) can

be expanded so that the header tapes and other outputs used in microfiche production can also be generated directly from the database.

In subsequent stages, the system can be expanded to support microfiche production and printing of full documents. The use of optical disk mass storage to complement microfiche generation is an area with high potential for favorable impact on DTIC users, as the full text of documents can be made available online.

3.3 DTIC Pilot System Components

Figure 5 shows an overview of the pilot system architecture. Individual workstations are monitored and/or controlled by the Master Document Control Subsystem (MDCS). The MDCS screens user and data access to the central library and monitors the document status as it moves from a pure image form, through character recognition, through the verification process, and through supplemental manual cataloging and indexing. Document input scanning and verification are accomplished at workstations. Other workstations are used to perform general purpose data entry and administrative functions.

3.3.1 Control Software

The control software, resident on a super microcomputer, will be CADEX, a commercial product of Database Applications, Inc. The combination of the super microcomputer, CADEX, and supplemental interface software forms the MDCS.

CADEX is an electronic card catalog. User–defined electronic cards are used to capture, store, and retrieve database information, and to point to documents which the cards may reference. Security features within CADEX regulate access to the documents and cards under its control, and a combination of its standard features provide full traceability for documents as they proceed through the system. CADEX offers a wide range of retrieval mechanisms for both experienced and inexperienced users. All card fields are fully indexed, and retrieval can be based on words or fragments within any field, or on keywords from a fully–linked broader and narrower term thesaurus. CADEX offers an efficient off–the–shelf solution to the MDCS control, indexing and retrieval requirements.

3.3.2 Workstations

Document input and modification will be accomplished at specially tailored workstations. The workstations are based on the IBM PC and use Database Applications' MicroCADEX to communicate with and exchange information with the MDCS. MicroCADEX provides the common user interface for each workstation. Underlying software for operation of scanners and verification editing of OCR'd documents is invoked automatically through this simple menu–oriented environment.

A scanning workstation will be used to capture page images. It will be tied directly to a flatbed image scanner through a high–speed communications link. The images

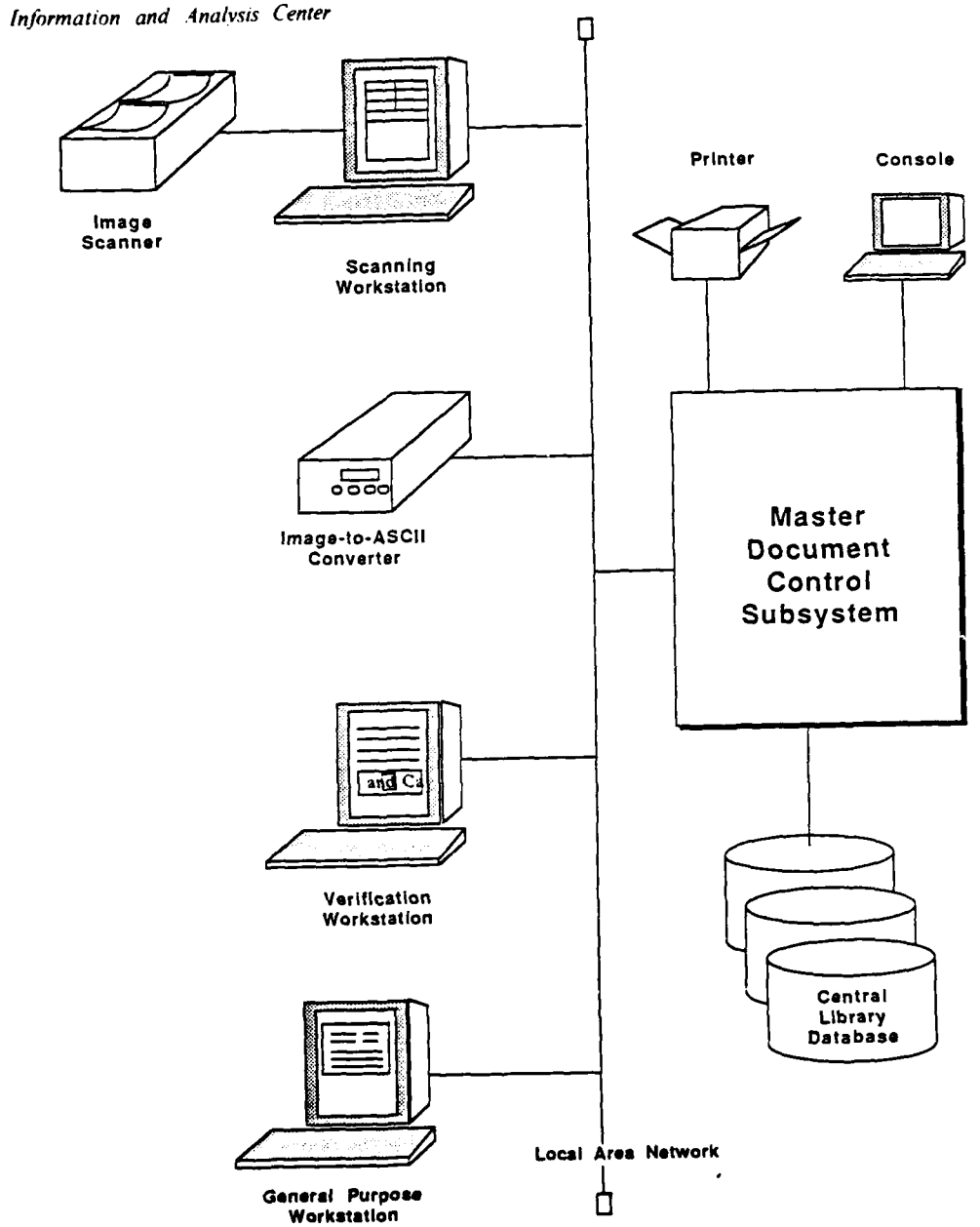


Figure 5: DTIC Pilot System Architecture

that are brought in at this workstation will be "matched" to the type of form being scanned, as illustrated in Figure 4, so that zones on the page can be cross referenced to the appropriate electronic card field. Anamet–supplied software will be used to interface the image scanner with the OCR device described in Section 3.3.3.

Following the OCR conversion, the proof editing pass will be performed on a verification workstation. Anamet–supplied software will be used to perform the verification operation and to exchange the resulting text information with the underlying database management/control software.

Other workstations will be less specialized in their orientation, and will provide general purpose tools for accessing the online information contained in the central library. Since the pilot system will address only textual information (no graphics), this workstation can have more limited software and hardware features than the scanning and verification workstations.

Workstations are connected through a local area network and can be physically located as needed in the DTIC work flow.

3.3.3 Optical Character Recognition

Optical character recognition is performed by a Recognition Server, a commercial product of the Palantir Corporation. The Recognition Server will be resident as a device on the local area network. CADEX tracks document images which enter the system at the scanning workstation and passes them to the Recognition Server for OCR. The unverified results from the Recognition Server are recaptured by CADEX for subsequent proofing at verification workstations.

3.3.4 Local Area Network (LAN)

An Ethernet LAN will be used for communication between the workstations and the MDCS, and for transferring document images and ASCII text. Ethernet communications is supported by the wide variety of hardware assembled for the pilot system, including the super microcomputer, workstations, and Recognition Server.

3.3.5 Image Processing

Image manipulation and enhancement are necessary to provide effective tools for dealing with real–world forms and documents. Completely software–based solutions to these functions and to image compression/decompression are generally too slow for the anticipated production requirements at DTIC. For this reason, the pilot system development effort includes integration of a hardware solution to these problems using a PC–based Raster Image Processor (RIP) board.

3.3.6 Storage

The initial pilot system will acquire and track forms–based information for subsequent entry into the TR database. Its storage capacity is dictated by the number of doc-

uments anticipated in a single update cycle at DTIC. Additionally, the information contained within the database is highly dynamic in nature, requiring verification, editing and additions. For these reasons, magnetic disk storage is recommended for the pilot system. The system, as designed, can accommodate optical storage media in concert with longer–term EDS goals.

4 TECHNOLOGY IMPLEMENTATION WITHIN DTIC DOCUMENT WORK FLOW

DTIC has tasked Anamet Laboratories with reviewing DTIC's current cataloging operations and assessing the impacts of the new technologies on them. The primary objective has been to identify impact areas within the existing operations and to improve throughput in those areas judged critical. Changes made to satisfy the primary objective must provide a clear migration path toward longer term DTIC EDS goals. The emphasis to date has been to limit the initial impact of the technology to those operations where its benefits are immediately needed, without affecting the daily activities of staff in other areas.

In its review, Anamet has utilized data developed by DTIC in past efforts in combination with interviews with cognizant DTIC personnel. DTIC studies which have proved particularly useful include those performed by the Logistics Management Institute [5,6] and Randy Bixby [7]. These studies formed a foundation on which to build a "straw man" model, which was then discussed at length with the DTIC personnel responsible for the key operational areas. The resulting flow diagrams represent, to our best understanding, a consensus of all parties regarding actual operations.

The emphasis presented here is on the processing of Technical Reports. TR's represent DTIC's primary workload by volume, and they pass through all processing stages shown in the diagrams. By contrast, DTIC271, DD1498, and PEDS entries are processed in small batches and are not subject to all the processing activities shown in the diagrams. From the process flow perspective, they can be thought of as a subset, with a TR–based comparison between current and projected processing being valid for them as well.

The diagrams focus on the comparison between current operations and those that would result with application of the candidate technologies. Documents and descriptive forms generally flow from left to right in the diagrams as they move through the system, and the light grey areas in the Stage I diagrams show the processing operations that are affected by the proposed technology.

4.1 Current Document Processing Operations

Figures 6 through 11 diagram the processing steps affecting Technical Reports (TR) submitted to DTIC for inclusion in the TR database. Figure 6 shows the document flow in overview and Figures 7 through 11 show operations within each of the major DTIC organizational branches.

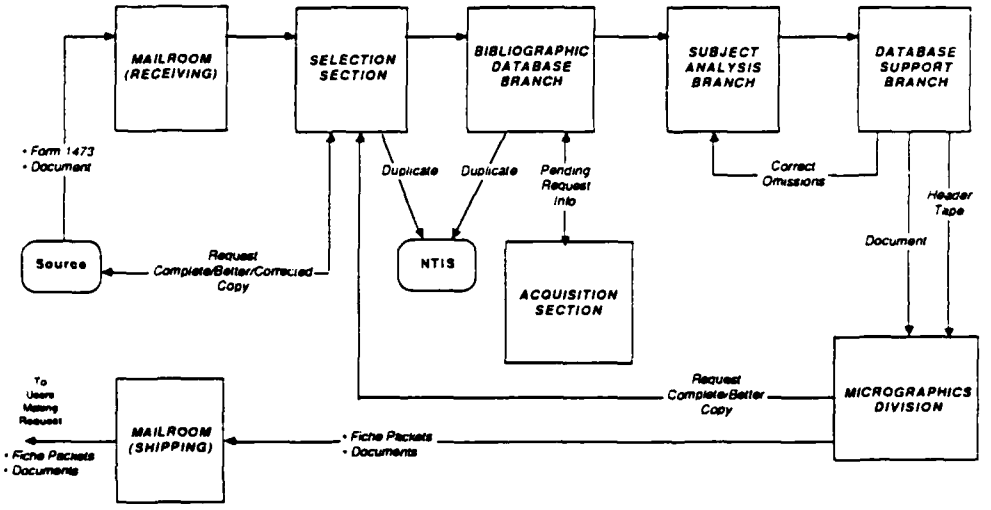


Figure 6: Current DTIC Work Flow

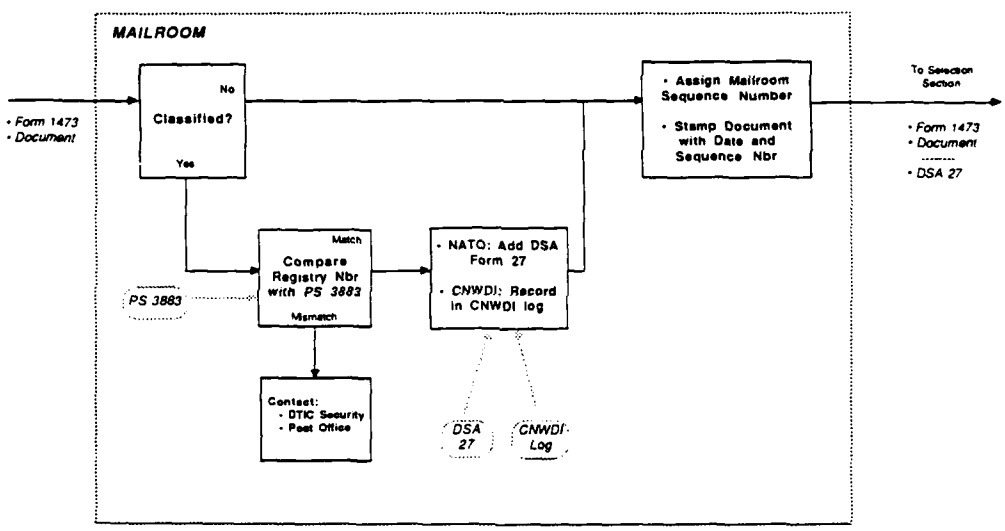


Figure 7: Current Mailroom Work Flow

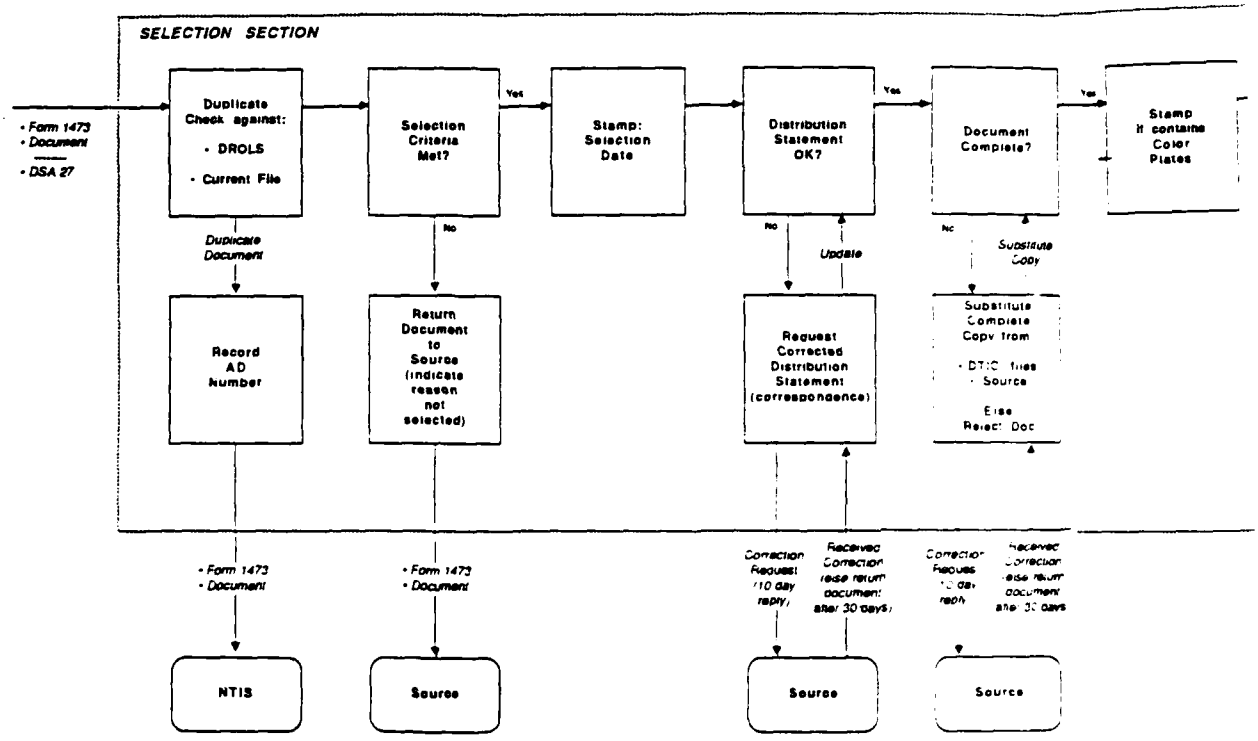
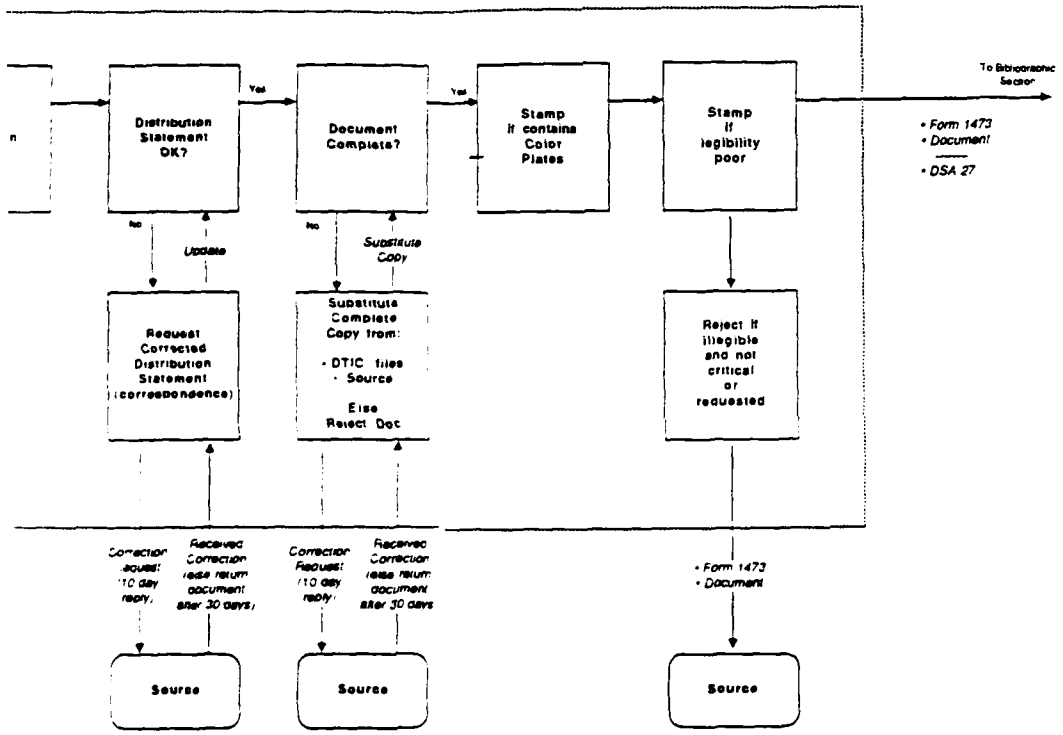


Figure 8: Current Selection Section Work Flow


Distribution Statement OK?
Yes
No
Update
Request Corrected Distribution Statement (correspondence)
Document Complete?
Yes
No
Substitute Copy
Substitute Complete Copy from:
- DTIC files
- Source
Else Reject Doc
Stamp If contains Color Plates
Stamp If legibility poor
Reject If illegible and not critical or requested
To Bibliographic Section
- Form 1473
- Document
- DSA 27
Correction Request (10 day reply)
Received Correction release return document after 30 days)
Correction Request (10 day reply)
Received Correction release return document after 30 days
- Form 1473
- Document
Source
Source
Source

Flow

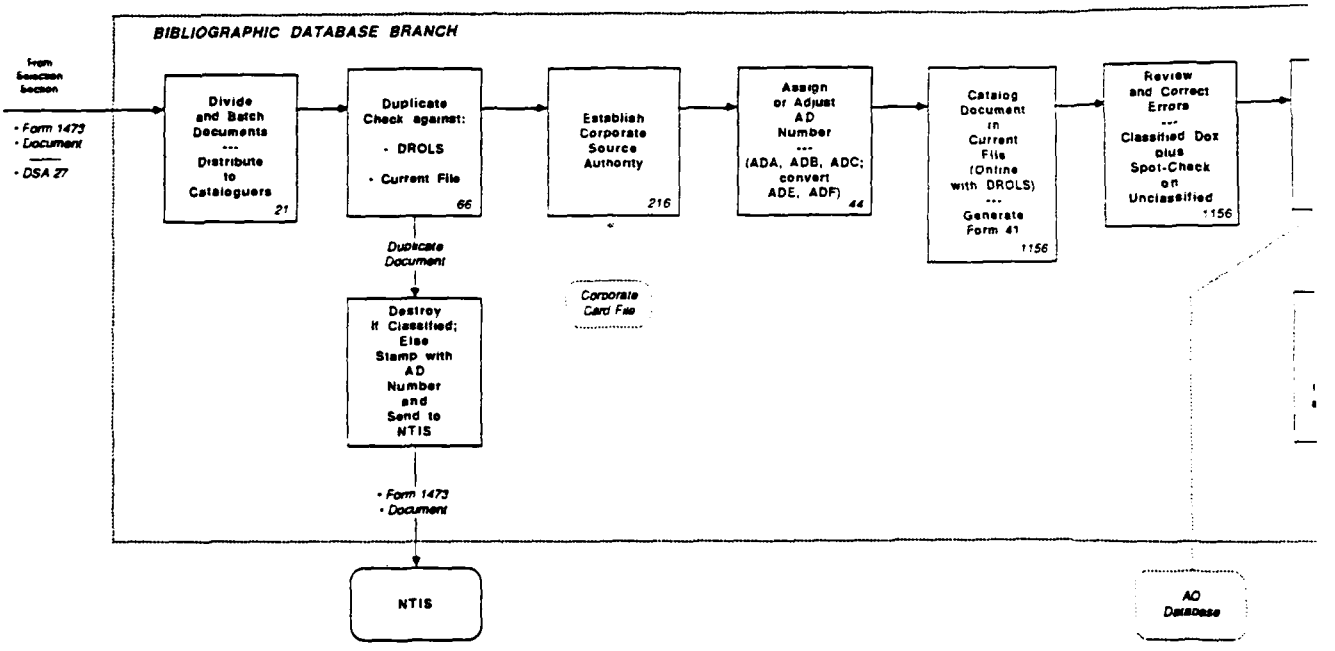
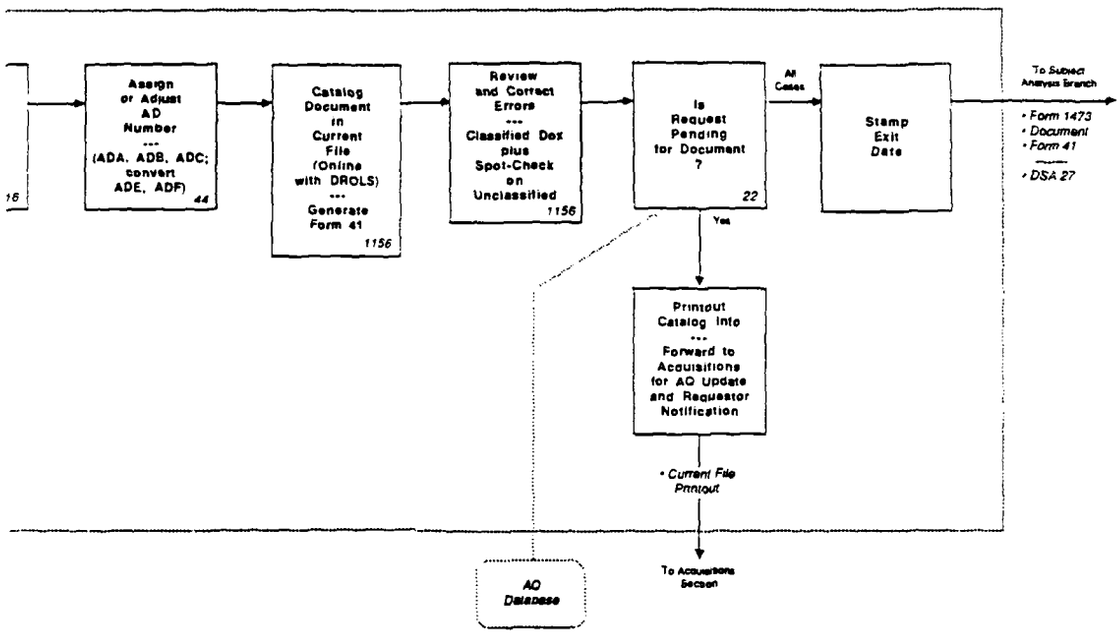


Figure 9: Current Bibliographic Database Branch Work Flow


Assign or Adjust AD Number --- (ADA, ADB, ADC; convert ADE, ADF) 44
Catalog Document in Current File (Online with DROLS) --- Generate Form 41 1156
Review and Correct Errors --- Classified Dox plus Spot-Check on Unclassified 1156
Is Request Pending for Document ? 22
Yes
All Cases
Stamp Exit Date
To Subject Analysis Branch
• Form 1473
• Document
• Form 41
• DSA 27
Printout Catalog Info --- Forward to Acquisitions for AO Update and Requestor Notification
• Current File Printout
To Acquisitions Section
AO Database

Work Flow

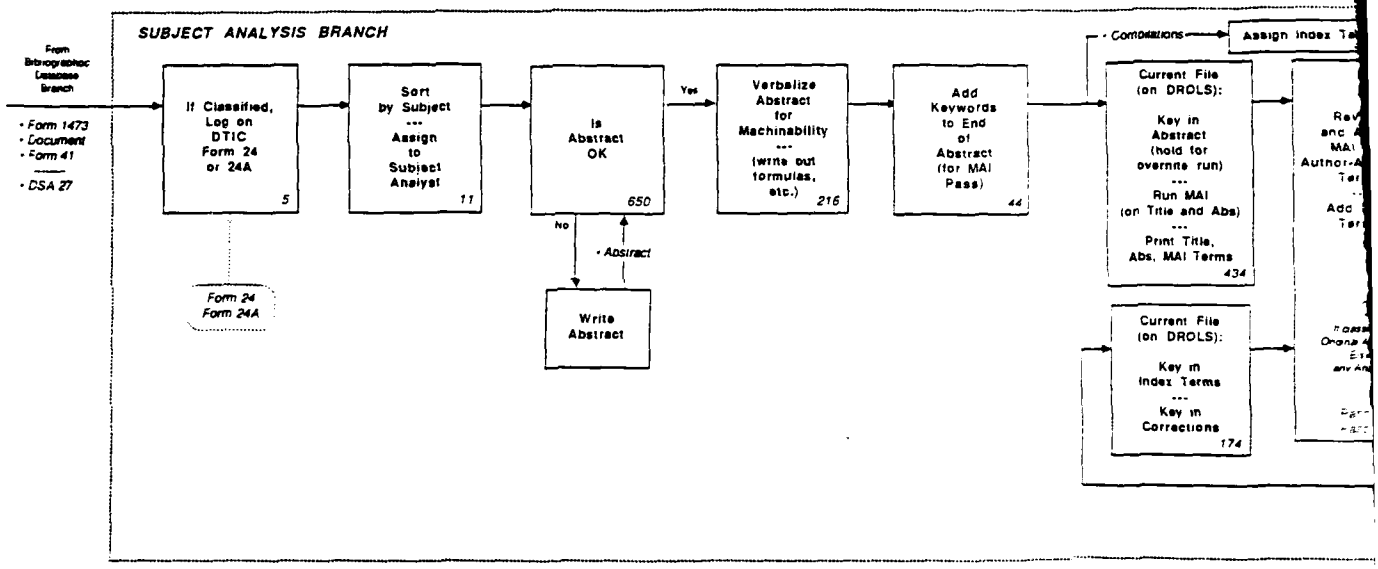
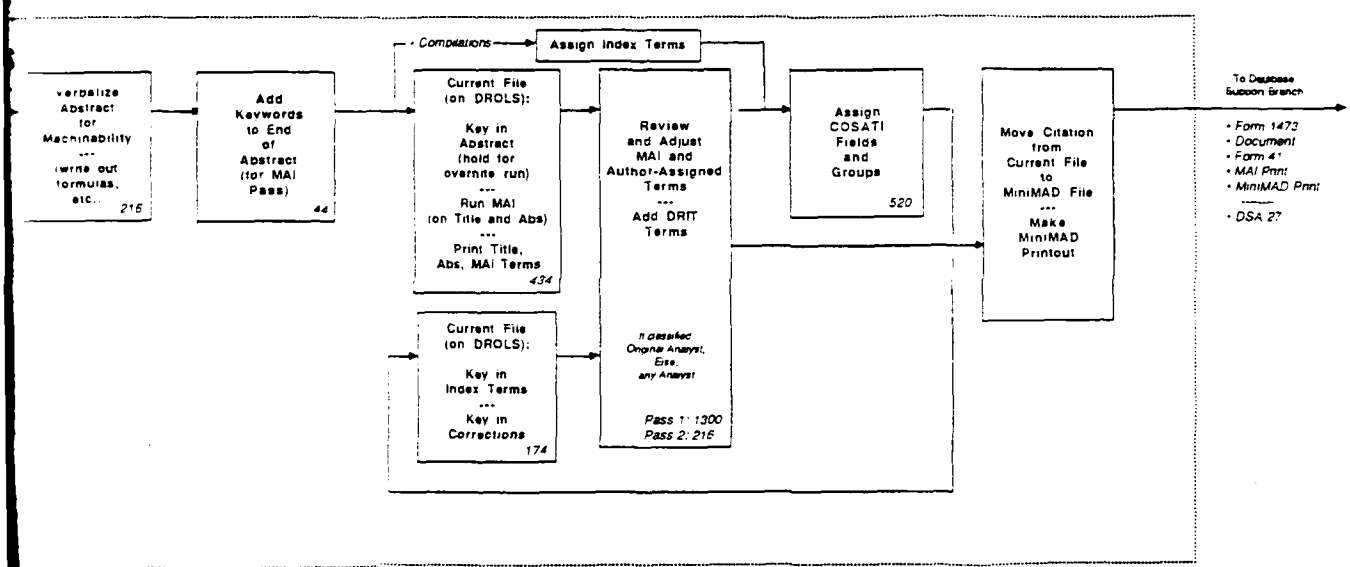


Figure 10: Current Subject Analysis Branch Work Flow


verbalize Abstract for Machinability

(write out formulas, etc...
216
Add Keywords to End of Abstract (for MAI Pass)
44
Comparisons
Assign Index Terms
Current File (on DROLS):
Key in Abstract (hold for overnite run)

Run MAI (on Title and Abs)

Print Title, Abs, MAI Terms
434
Current File (on DROLS):
Key in Index Terms

Key in Corrections
174
Review and Adjust MAI and Author-Assigned Terms

Add DRIT Terms
If classified: Original Analyst, Else, any Analyst
Pass 1: 1300
Pass 2: 216
Assign COSATI Fields and Groups
520
Move Citation from Current File to MiniMAD File

Make MiniMAD Printout
To Database Support Branch
• Form 1473
• Document
• Form 41
• MAI Print
• MiniMAD Print
• DSA 27

OW

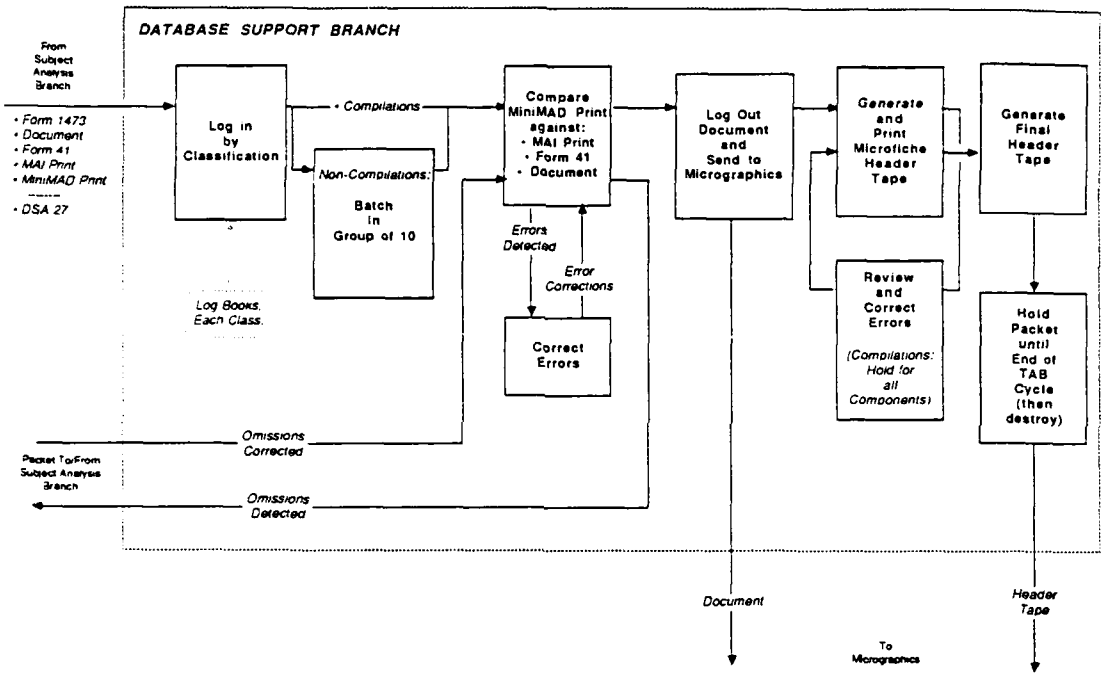


Figure 11: Current Database Support Branch Work Flow

In overview (Figure 6), documents enter the system through DTIC's Mailroom, along with the DD1473 or other forms describing their content. Information from the DD1473, as supplied by the source of the document, provides the foundation for the TR data entry. As DTIC's cataloging personnel review the document, they supplement and adjust the information from the DD1473 to formulate the record that will eventually become part of the TR database. The document flows with the developing record throughout processing to serve as a reference, and is sent to the Micrographics Division to be recorded on microfiche after all data analysis has been completed. Each microfiche carries header information identifying it with its TR database record, and this header information is supplied on tape to Micrographics for use during filming. The microfiched document becomes part of the permanent film library, and the TR database record for it makes it accessible to users. During initial micrographic processing, microfiche copies are made for users participating in the Automatic Document Distribution (ADD) program.

Looking inside each of the major divisional boundaries (Figures 7 through 11), operating details relevent to scanning technology application become visible.

4.1.1 Mailroom (Receiving)

In the Mailroom, each document receives a stamped receipt date and sequence number. While the sequence number is not used by the other DTIC branches, it allows the Mailroom to associate duplicate copies of the document (which it has retained during processing) with the copy that has been sent through for processing. The permanent accession number under which the document will be filed in the TR database is not assigned until selection and various categorization decisions are made.

4.1.2 Selection Section

Selection Section personnel decide if the document meets requirements for inclusion in the TR database. Because deficiencies in the document's legibility or distribution list may have to be remedied by the document's supplier, documents can be held for as long as a month while correspondence with the source is exchanged. Documents not in accord with DTIC's mission or selection criteria may also be rejected at this stage, and a cursory duplicate check is performed to identify documents already in the database. Documents that exhibit no deficiencies can flow through the Selection Section's activities in one day, but those that are held may remain for as long as a month. Since the recordation of the document in the Current File (i.e., as having been received and being in process) occurs downstream in the Bibliographic Section, a "lost month" of traceability has become a problem. This lost month particularly impairs the notification of users who entered a request for the document and who may have an urgent need for it.

4.1.3 Bibliographic Database Branch

After selection, documents and their DD1473 forms are sent to the Bibliographic Database Branch. A thorough duplicate check is performed against both DROLS and the Current File, based on cataloging information such as author, performing organization, date, title and report type. By contrast to the duplicate check performed upstream in the Selection Section, this check is a primary filter for the TR database and is performed by cataloging personnel. Determining whether a document already has an entry in the database is often not simple. For example, the only difference between an interim report and a final report for the same project may be a report type code, with the source sending in a photocopy of the DD1473 from the interim report when submitting the final.

Next, the corporate source of the document is checked against a manual card file maintained by the Bibliographic Database Branch, resolving such issues as corporate name changes that may have occurred. For a DROLS user trying to trace development of a corporate technical expertise in some area, unambiguous treatment of these issues is important.

The document is then assigned a permanent accession number from one of five series (AD–A, B, C, D (patent), or P (compilations)), depending on its distribution restrictions (unclassified, limited, and classified, or special document types). Documents submitted electronically by one of the Shared Bibliographic Information Network (SBIN) sites have their temporary AD–E and AD–F numbers converted to one of the above series for permanent storage.

The entire range of catalog information from the DD1473, less the abstract field, is then keyed into the Current File via remote terminal access. The keyed input is verified for all classified documents, but workload constraints prevent verification of unclassified documents except on a spot–check basis. The personnel performing keyed entry and verification are trained catalogers, and their cognitive knowledge of DROLS and the technical data they are entering helps them to identify errors. The keyed entry is nonetheless fatiguing, as is the existing verification process. Verification involves comparison of the DD1473 original against a printout of the entered data.

Lastly, the Bibliographic Database Branch reviews the separate Acquisitions (AQ) Database to determine if a request has been made for the document. If so, an additional copy of the catalog information is made and forwarded to the Acquisitions Branch so that they can notify the requestor that the document has arrived and can update the AQ database.

In Figure 9, the action blocks on the main processing chain within the Bibliographic Database Branch contain numbers in their lower right corners indicating the number of hours per TRAC (Technical Report Awareness Circular) cycle devoted to the activity. For each block, the number shown is the throughput (in documents per hour) divided into the 2,600 documents processed during the average monthly TRAC cycle. The throughput values used in these figures were reported for the activity either by DTIC personnel or by previous DTIC–sponsored studies. The person–hours per TRAC presentation shows clearly that DTIC's perception that keyed entry is the primary bot-

tleneck can be borne out quantitatively. Between the 1,156 person–hours per TRAC spent in actual keyed entry, and the 416 person–hours per TRAC devoted to review and correction, 81 percent of the Bibliographic Database Branch's document processing time is accounted for.

4.1.4 Subject Analysis Branch

From the Bibliographic Database Branch, documents flow to the Subject Analysis Branch, where additional cognitive analysis is performed. The supplied abstract (if any) is scrutinized. Adjustments may be made to make it more meaningful to DROLS users or to improve the results of the subsequent machine–aided indexing (MAI) pass. Equations and other non–standard character insertions are "verbalized" into word equivalents so that the MAI pass will pick them up. If the analysts judge that the supplied abstract does not provide a description that will be meaningful to DROLS users, or if no abstract is supplied, they either write an abstract or construct one by highlighting and tying together sentences from other portions of the document. Index terminology that the analyst feels should be included in the MAI record is appended to the end of the abstract so that the MAI software will pick it up.

All these operations are performed on paper by marking up copies of the DD1473, or the appropriate document sections, or by notation on the Form 41. At completion the marked–up abstract is passed to word processing personnel, who key it into the Current File and queue it for an overnight MAI run. The MAI output, delivered for analyst review the next day, contains the keyword terminology covering the title and abstract as suggested by the MAI software, along with the title and abstract text, once again in hardcopy. These outputs are reviewed along with the original document package and are marked up with any corrections. The subject analyst then assigns controlled vocabulary terms from the DTIC Retrieval and Indexing Terminology (DRIT) and COSATI field and group codes appropriate both to "need to know" distribution restrictions that may be applicable and to general searching. The hardcopy document package then goes back to the word processing personnel for entry of the additional information and correction of errors that have been detected.

After the second data entry pass, a corrected Form 41 is generated, added to the document package and returned to the analyst for review. If further corrections are noted, the package cycles one more time. If it is error–free, the document's record in the Current File is released for inclusion in the "MiniMAD" file, which is loaded into the TR Master Accessioned Document (MAD) file at completion of the TRAC cycle. A hardcopy of the MiniMAD entry is generated and added to the document package, which is then passed on to the Database Support Branch.

As in the Bibliographic Database Branch, data entry operations have become a key bottleneck in the Subject Analysis activities. The cumulative person–hours show the review and verification activities to be more time–consuming than the actual data entry, but the repeated cycling takes its toll both in data entry and in repeated reviews of the same document package. Additionally, while the initial paper–based abstract review is not in itself more labor intensive than it would be online, the markup, correction,

verbalization and other operations may well be. A fraction of the time utilized in these activities probably represents an indirect saving area, if the staff involved can be provided with online access to the document record. The bottom line in current DTIC operations, in any event, is that because of the repeated passes through data entry, document packets awaiting either initial entry or correction tend to converge on the data entry personnel late in the TRAC cycle, becoming a key pacing factor in the overall operation.

4.1.5 Database Support Branch

After Subject Analysis, the document packet is sent to Database Support for a final check and for generation of the header information (on computer tape). Headers are reproduced onto the image of each microfiche page when Micrographics subsequently films the document. While Database Support has direct access to the MiniMAD entry and can correct errors online, it must cycle the packet back to Subject Analysis if omissions are detected, for correction by the subject analysts. This flow, again paper-based, is an additional convergence path on the data entry personnel in Subject Analysis. Documents approved for release are sent to Micrographics for filming, and their header information follows the next day.

The final catalog information check performed by Database Support, while not a major time consumer in itself (6 documents/hour, or 217 person-hours per TRAC cycle), is significant in that it represents a recheck of information that has previously been entered and checked against a machine-generated output format. That is, the MiniMAD data should be directly derivative from the Form 41 and other data against which it is being checked. Since errors and omissions are picked up at this stage, this check is currently needed; however, if the accuracy and completeness of the upstream checks can be improved, then the automatically-generated nature of producing the MiniMAD data can be exploited to save this effort.

4.2 Proposed Stage I Operations

The intent in Stage I is to apply scanning and database technology selectively to reduce some of the key document throughput bottlenecks, specifically those in the data entry and verification area. Summarizing the operations in the Bibliographic Database and Subject Analysis Branches directly associated with catalog data entry from the DD1473 and its verification, approximately fifty percent of the operations of these two branches are affected by this step. The intent in Stage I is to retain all other operations in their present form at the outset. Those operations currently performed on paper will remain so.

Entry of the DD1473 data via scanning and OCR, and automatic collection of the data into a directly coupled central database is expected to have two immediate effects. First, the time-consuming keyed entry operation is supplanted by scanning and OCR: data entry becomes very similar to feeding book pages to an office copier, and the OCR-conversion, which is more time-consuming, is automatically scheduled and

proceeds without operator intervention. Second, the verification needed to assure that the data automatically loaded into the online database is error–free can be provided with a single, computer–aided correction pass through the converted data.

The verification mechanism described in Section 3.1 provides a much less fatiguing and more effective approach than the present keyed–entry and verification cycle in use at DTIC. It is anticipated that a simple "read through" will be all that is needed as a final confirming check, with the need eliminated to laboriously compare every phrase of the original against what was produced. The result should be reliable, error–free data, produced very close in the processing flow to the original data entry point.

Image scanning operates at 2 to 10 seconds per form page and verification can be expected to take perhaps 2 to 5 minutes per form page by these methods. OCR conversion, which may take 1 to 2 minutes per page, is an unattended operation. It should therefore be possible to make major improvements in staff utilization in these tedious and error–prone portions of the processing operation, without upsetting the functions of any other workers. Outputs to paper will produce the Form 41 and other packet components currently in use, and outputs to computer media will produce the feeder information for the MiniMAD file.

In overview (Figure 12), DTIC operations look very much as before, but certain of the major divisions ("grey" areas in the figure) have internal changes in their operations, as shown in Figures 13 through 15. From the overall flow perspective, the major change is that scanning, OCR, and verification activities, operated by the personnel in the Database Support Branch currently responsible for keyed entry, are inserted after Selection, but before the document goes to the Bibliographic Database Branch for cataloging. In these activities, the entire DD1473 (abstract as well as catalog information) is scanned, converted, verified and loaded into the central library database. The data entry and verification functions previously performed by the Bibliographic Database and Subject Analysis Branches become editing and reviewing functions on data that is already online. The review becomes a cognitive one, rather than a character by character scrutinization. It should therefore be dramatically faster and much better matched to the skills of the cataloging personnel.

Additionally, it should be possible to eliminate the "lost month" of traceability that has been identified by DTIC personnel as a problem. In current operations, no record is made of a document's arrival at DTIC until after selection. There is therefore no easy way of advising someone who has ordered the document that it has arrived until selection problems have been corrected. Because of the outside correspondence involved, this process may take up to a month and is not within DTIC's control. A month may therefore be lost before the document becomes visible.

Documents with no selection problems can get through selection in one day, and only those with problems to be resolved are held. By a slightly modified routing, the "lost month" can be eliminated. Specifically, as shown in Figure 12, all documents are passed on for scanning. Those with no selection problems proceed into the cataloging activities, and those which need additional selection treatment are returned to Selection. The Current File database has then received an input event for the document very close to

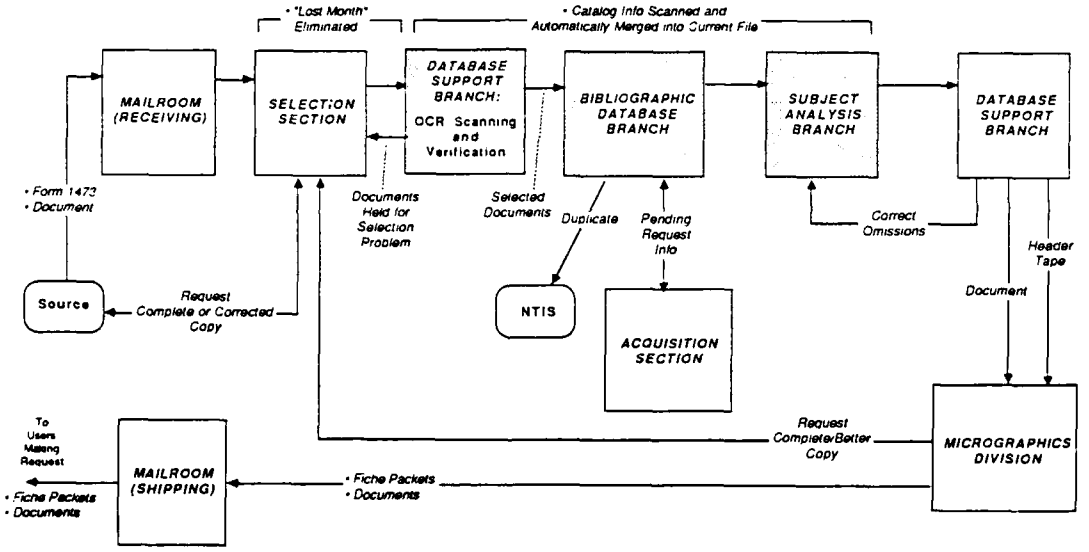


Figure 12: Stage I DTIC Work Flow

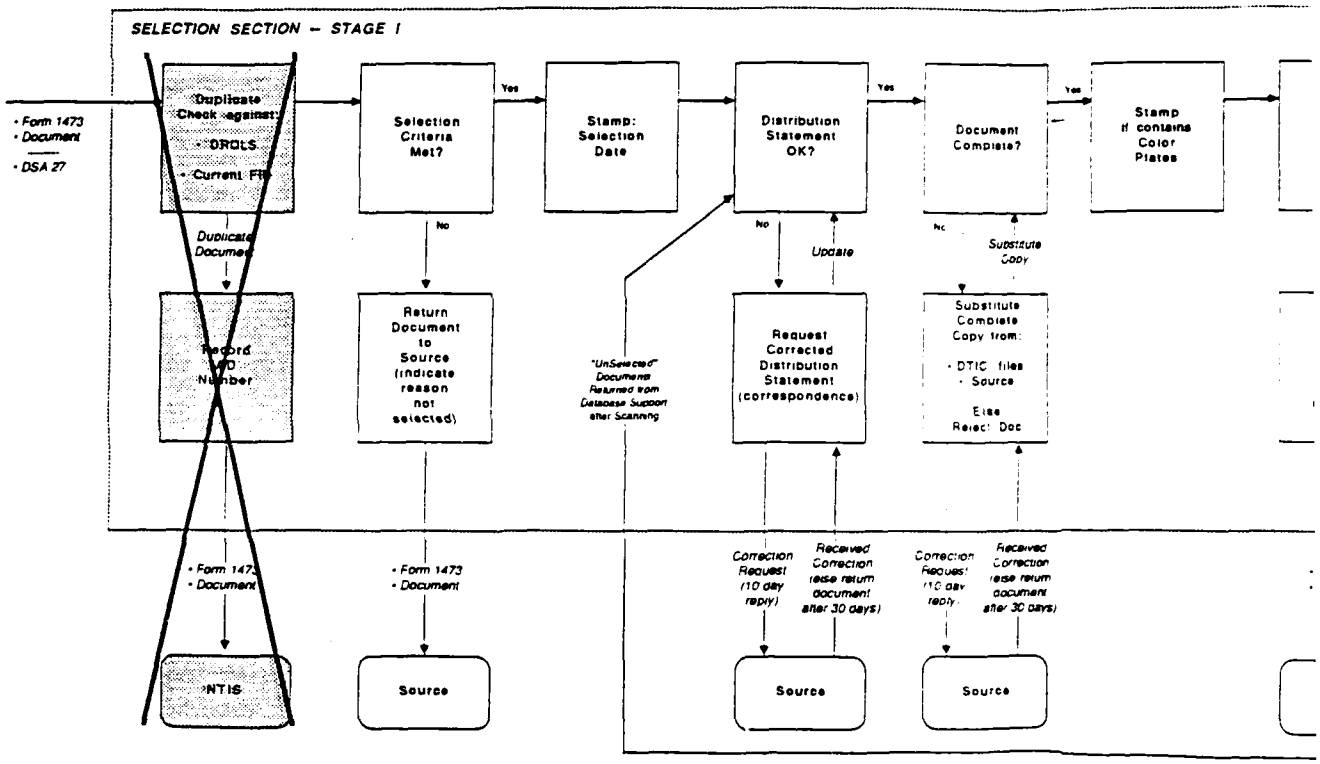
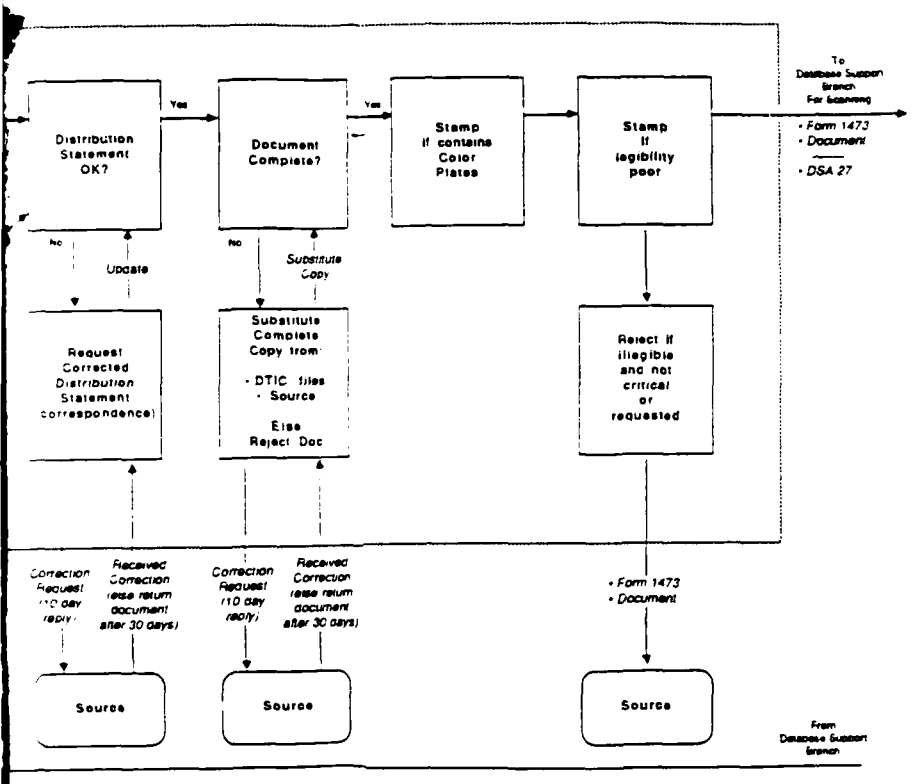


Figure 13: Stage I Selection Section Work Flow


Distribution Statement OK?
Yes
No
Update
Request Corrected Distribution Statement correspondence)
Document Complete?
Yes
No
Substitute Copy
Substitute Complete Copy from:
- DTIC files
- Source
Else Reject Doc
Stamp if contains Color Plates
Stamp if legibility poor
Reject if illegible and not critical or requested
To Database Support Branch For Scanning
- Form 1473
- Document
- DSA 27
- Form 1473
- Document
Correction Request (10 day reply)
Received Correction (else return document after 30 days)
Correction Request (10 day reply)
Received Correction (else return document after 30 days)
Source
Source
Source
From Database Support Branch

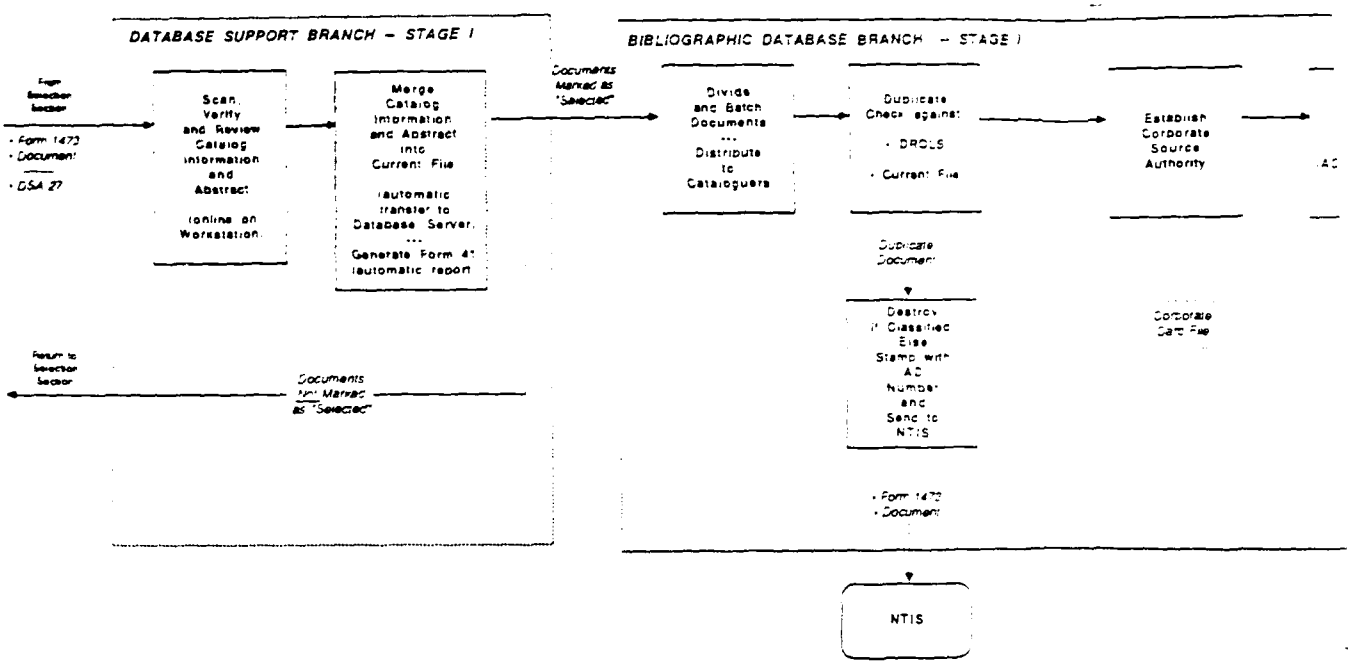
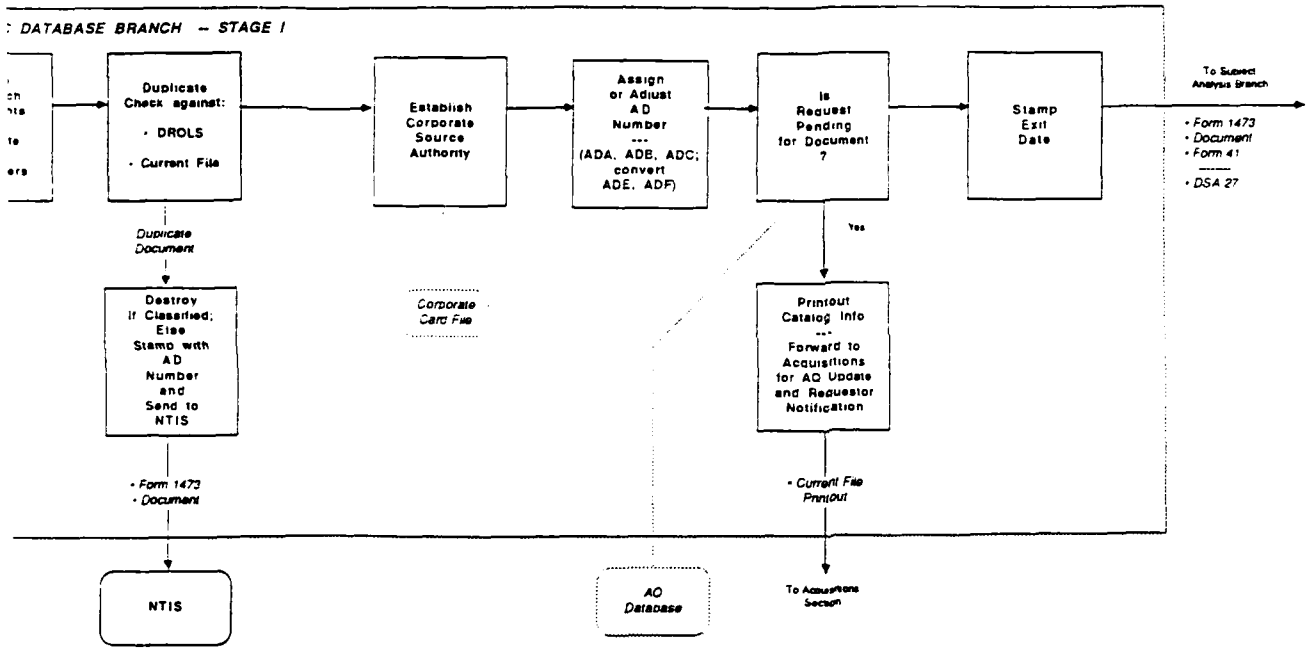


Figure 14: Stage I Database Support and Bibliographic Database Branch Work Flow



ıch Work Flow

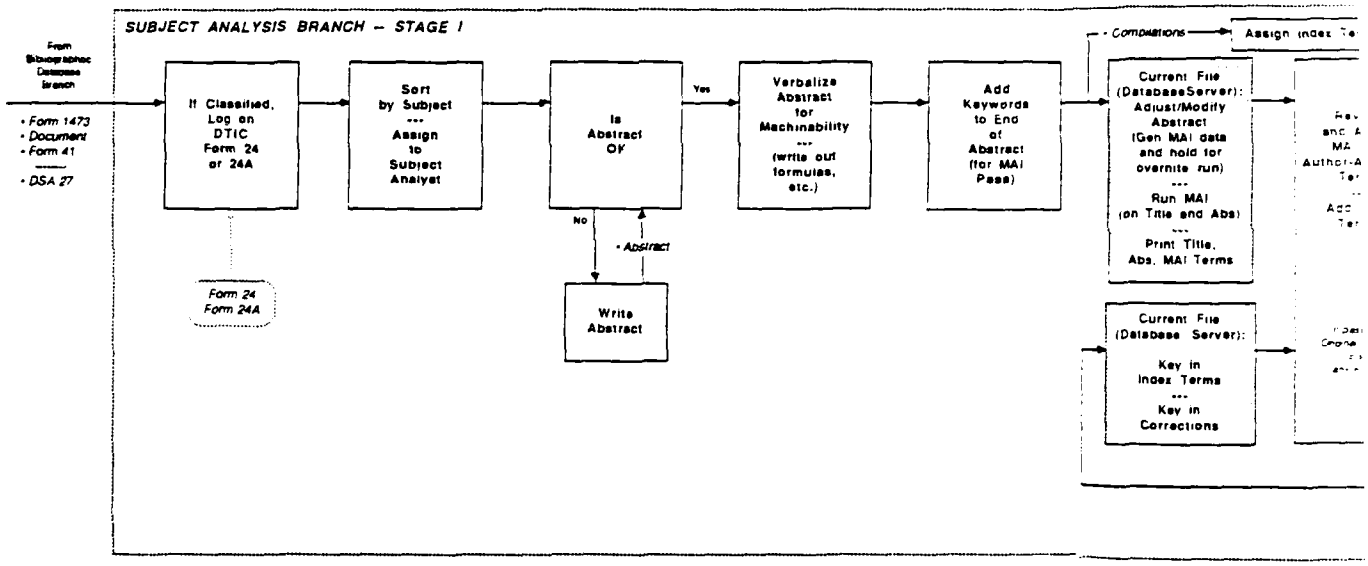
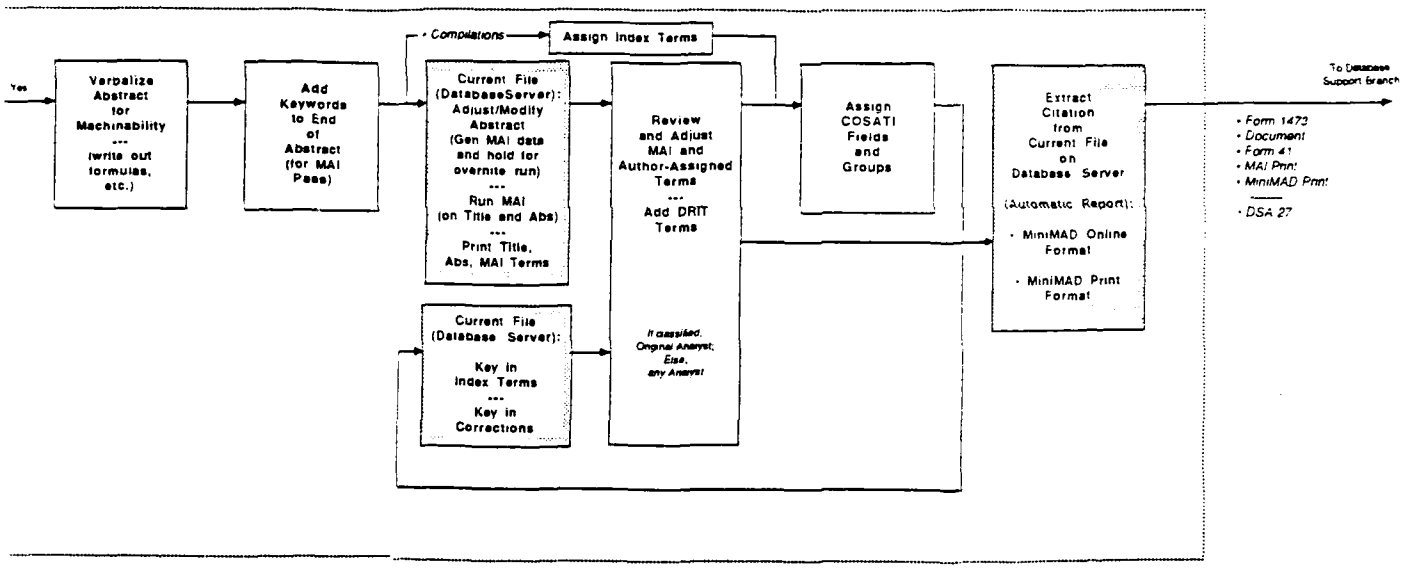


Figure 15: Stage I Subject Analysis Branch Work Flow


Yes
Verbalize Abstract for Machinability

(write out formulas, etc.)
Add Keywords to End of Abstract (for MAI Pass)
• Compilations
Assign Index Terms
Current File (DatabaseServer):
Adjust/Modify Abstract
(Gen MAI data and hold for overnite run)

Run MAI (on Title and Abs)

Print Title, Abs, MAI Terms
Current File (Database Server):
Key in Index Terms

Key in Corrections
Review and Adjust MAI and Author-Assigned Terms

Add DRIT Terms
If classified, Original Analyst; Else, any Analyst
Assign COSATI Fields and Groups
Extract Citation from Current File on Database Server
(Automatic Report):
• MiniMAD Online Format
• MiniMAD Print Format
To Database Support Branch
• Form 1473
• Document
• Form 41
• MAI Prnt
• MiniMAD Prnt
• DSA 27

·k Flow

its original arrival time at DTIC, and the document's status can be queried at any time thereafter. For documents that are not eventually selected for inclusion in DROLS, the only additional labor is their scanning and verification, which is not major because of the computer–aided methods, and minor additional physical movement of the document among divisions. This appears to be a small penalty for the major improvement in traceability, and is therefore a recommended step.

In the initial implementation under Stage I, no attempt will be made to provide direct scanning and OCR support for the abstract synthesis operations that are performed by the Subject Analysis Branch when the abstract supplied by the source is inadequate. However, the software components being developed for use in associating zones on the DD1473 with the internal database fields into which the data is to be loaded can be slightly reconfigured to serve this purpose. In Stage I, then, abstract synthesis and verbalization are left as at present. It is anticipated, however, that a short term follow–on effort can support scanning and OCR entry of sentences from marked up sections of the actual document, and the automatic appending of these into a synthesized abstract. As in the initial Stage I operation, this abstract can then be edited as desired.

4.3 Proposed Stage II Operations

Stage II is anticipated to be a gradual extension of the Stage I system to support additional operations within the DTIC processing procedure. Steps intentionally left on paper in the initial implementation will be brought online one at a time, as workstations are provided for the members of each activity. The abstract synthesis function described above is an example of such a functional extension.

Another extension might be the movement to online operation of the keyword indexing activities. This extension would include direct access by the indexing personnel to the Current File database, so that they could make additions online. In working with the controlled vocabulary (DRIT), online operation would allow provision both of direct validation against the thesaurus of keywords proposed for use on a particular document's record, as well as the browsing of the DRIT to find the optimum terminology to apply. Both these enhancements make use of existing features in the CADEX database software being used in the prototype system, and they can be provided without major effort to meet the compatibility requirement between the terms of the DRIT vocabulary and the TR database.

A further addition would be the automatic validation of field information against restricted value sets as the information is added to the database. Fields containing cognitive errors would be flagged for correction, much as the computer–aided text verification cited for Stage I helps to focus effort on the text conversion errors. A number of DTIC personnel have indicated that such functionality would be of great utility in their operations.

Generally, the tone of Stage II is to bridge between the sections initially operating online, filling in gaps in a progressive manner. Additionally, the Current File focus will

be extended back into the operations of the Database Support Branch to encompass those activities dependent on the MiniMAD file, such as generation of the microfiche header tape.

5 CONCLUSIONS

The use of integrated OCR and database management technology to improve DTIC document input processing has been examined. Significant near–term improvements in efficiency can be realized using commercially available components fused into an integrated system. An approach and system architecture have been defined that will permit a staged implementation of this technology within the framework of the current DTIC work flow. While the emphasis in this effort is on reducing labor–intensive manual keystroking operations presently in use, the proposed system provides an open ended approach which will interface easily with both existing and future DTIC operations.

An in–depth review of current document processing work flow was used to guide the definition of the pilot system architecture developed in this study. The initial study and implementation recommendations were presented to cognizant DTIC personnel for review, and the revised work flow definition presented in this report reflects a consensus viewpoint.

A demonstration system is scheduled for DTIC review near the end of the present fiscal year, followed by hands–on operation of the system at DTIC in the production environment.

References

[1] Saunders, J., "DTIC Electronic Document System," February 1987.

[2] "MAIRS Requirements and Status Report," Anamet Report No. 86.015, April 1986.

[3] "Selection Process for a Document Storage and Retrieval System on MAIRS," Anamet Report No. 85.057, November 1985.

[4] "Technical Proposal for the Development of a MIL–STD Automated Indexing and Retrieval (MAIR) System," Anamet Report No. 85.001, January 1985.

[5] Egan, D., et al, "Functional Description for a Replacement Input System," LMI, AD–A174522, July 1986.

[6] Egan, D. and Travers, M., "A Baseline Description of DTIC Scientific and Technical Information Support Systems," LMI, September 1986.

[7] Bixby, R., "Current Document Handling Procedures at Defense Technical Information Center," AD–A162–200, November 1985.